Flyber Data Strategy MVP

Introduction

Flyber has been massively successful. Results have beaten expectations and projections! This is good news for Flyber, but now it's time to plan for what's next. With success came some challenges. While we were able to grow, the original data pipelines to receive and process data are unable to keep up with the current and future .

As a Data Product Manager, working with multiple teams and stakeholders is imperative to success. To understand what our needs are, what scale we are growing at, and how we can build for the future, we need to consider all relevant stakeholders. In this proposal, present your findings along with the analysis and reasoning behind the choices made in order to help Flyber continue its success.

Section 1: Data Customers & Needs

Flyber is a two-sided platform. You have **customers** who are **riders**, and you have **partners** who are **drivers/pilots** (think Uber: riders and drivers). For the **Minimum Viable Product, you will be focusing on the Riders side of the business**. To build an end to end data pipeline the very first step is to understand who needs data and why they need that data. Within Flyber, identify who your primary data customers/stakeholders are, why they are your primary data stakeholders and how they want to use the data (primary use-cases).

Identify your primary internal stakeholders and their use-cases:
(You may add more rows if necessary.)

Stakeholder	**Why are they primary stakeholders?**	**Use-Case**
Drivers	They are the main partner of Flyber, as they are the ones that interact directly with customers.	Business Intelligence
Marketing team	They need data for customer acquisition and strategy definition.	Business Intelligence Visualization

Finance team	Needs data in order to analyze product viability and growth.	Reporting
Engineering team	Needs data in order to make improvements to the product, in this case, the app/website.	Machine Learning
Product Management	Needs data to understand the customers' behavior and for real life monitoring.	Visualization Business Intelligence
Customer care team	Needs data to improve the reported issues and pain points.	Reporting

Section 2: Data Collection and Data Modelling

To support our primary stakeholders's use-cases we need following data:
(You may add more rows if necessary.)

Stakeholder	**Use-Case**	**Data**	**Why is this the primary use-case?**
Drivers (partner)	Business Intelligence	- Data about service usage trends, with information regarding the rides.	- Because it will help to understand the customers in more effective way and the drivers are the closest interaction stage with the customers.
Marketing team	Business Intelligence Visualization	- Data to monitor business trends. - Data to track the defined metrics.	- Because the Marketing team will use the data and the generated reports to identify the business trends and try to use the to acquire and retain customers.

Finance team	Reporting	- Data to analyze the financial status of the business.	- Because profitability, viability and business evolution should be monitored in order to make business modifications and to define strategies.
Engineering team	Machine Learning	- Data to personalize customer experiences (app/web).	- Because a good app/website will improve the conversion rates and will also be very important in order to gather customer data (entity and event data).
Product Management	Visualization Business Intelligence	- Data to identify user´s pain points.	- Because product managers need to identify where an improvement or a change is required and need reasons to convince the management.
Customer care team	Reporting	- Data will be used to provide personalized solutions.	- Because this information will be required to generate a history of the claims and will be valuable to identify patterns, so improvements can be done.

The tables we need are:

Note: As a best practice, we should establish these relationships between tables from the very beginning. To complete this exercise, we will focus on fundamental concepts of relational databases - tables, normalization and unique keys. Please provide the table header row for each table, tables might be different lengths. Make sure you include the following for each table. You can create as many tables as you feel are necessary (copy and paste from one of the table sections):

Table 1:

ride_table

Primary key □ ride_id	*Foreign key □ driver_id*
	Foreign key □ customer_id

Rationale for Choosing Primary and Foreign Keys for the Table 1:

In order to link all the information regarding the rides, we will use *"ride_id"* as the unique identifier (primary key). When data about the drivers is needed will link this table with the *"drivers_table"* through the *driver_id* field and we will use the *customer_id* field in order to retrieve the *customer_id* data when required.

Table 2:

purchase_table

Primary key □ purchase_id	*Foreign Key □ ride_id*
	Foreign Key □ customer_id

Rationale for Choosing Primary and Foreign Keys for the Table 2:

Every purchase will have data regarding the price of the transaction, the *customer_id*, the *ride_id* and more information about the purchase. The *purchase_id* will be the unique identifier for the purchases and the *ride_id* and the *customer_id* will be the fields that will enable retrieving the information regarding the ride (driver, timestamp, etc) and the customer (age, gender, etc.).

Table 3:

usage_table

(You may add more columns if necessary.)

Primary_key □ visit_id	*Foreign Key □ customer_id*

Rationale for Choosing Primary and Foreign Keys for the Table 3:

This will be an event table for the customer visit to the website or the app. Here the fields defined to describe the customer interaction will be registered. Every season will have a *visit_id* and will be connected to customer data through the *customer_id*.

Table 4 :

customer_table

(You may add more columns if necessary.)

Primary_key □ customer_id	

Rationale for Choosing Primary and Foreign Keys for the Table 4:

In customer table we will have just a primary key. In this table all the customer' attributes will be contained.

Table 5 :

driever_table

(You may add more columns if necessary.)

Primary_key □ driver_id	

Rationale for Choosing Primary and Foreign Keys for the Table 5:

In customer table we will have just a primary key. In this table all the drivers' attributes will be contained.

Table 6:

customer_rating_table

(You may add more columns if necessary.)

Primary_key □ rating_id	*Foreign Key □ customer_id*

Rationale for Choosing Primary and Foreign Keys for the Table 6:

The *rating_id* will be the primary key of this table because this will be the unique identifier of the event, hence, the event of the rating itself. The foreign key will be the *ride_id* so it is possible to track ride that is being object of the rating.

Section 3: Extraction and Transformation

Now that you have the requirements from your stakeholders, you want to understand the current state of what data is collected. That is how you recognize which additional data you need to achieve the future state. You ask the engineering team what data they are currently collecting in the pipelines and they provide you with section_3_event_logs template (which you can download from the classroom) generated by rider's activities on the Flyber App. Also provided in the Project Resources.

Extraction and Transformation-1

ETL is performed on the provided Event Logs Template and results will be transferred to the proposal template. The project's ETL should be created inside of your copy of the Event Logs template in the tab titled, ETL. Clicking on the **link above** will create a copy of the Event Logs for you

After being provided with a CSV log file, use extraction techniques to be able to get the data into a usable form. Because this needs to be a repeatable process we need to document it in order to assess its feasibility. Below,

1. Write the steps you took to extract the data and provide reasoning for why you used this method *Note: Don't forget to include any file type changes*:
2. Perform cleaning and transformation of the data in the ETL tab and document.
3. Document and provide rationale for all of your steps below as well.

Steps for Extraction:

1. *Convert "event_time"* using *"text to column"* from excel.
    a. Convert the column from date to text in order to extract the year, month and day.

2. *Convert the "event time" column to "time".*
    a. This was done in order to keep the time of the event.
3. Rename the column with the date to *"event_date".*
    a. In order to differentiate this column to *"event_time".*
4. Rename the columns with year, month and day with the same name.
    a. Correct column naming in order to make the information clear.

Transformation-2

Analyze the data from part 1 to answer the following questions:

1. How many events are being recorded per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Event Count	9,891	18,056	18,202	17,963	17,600	17,694	17,595

2. How many events of each event type per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Choose Car	1,498	2,843	2,953	2,769	2,725	2,801	2,804
Search	1,484	2,891	2,824	2,899	2,749	2,904	2,821
Open	6,594	11,733	11,767	11,662	11,531	11,325	11,371
Begin Ride	38	49	62	86	57	57	78
Request Car	277	540	596	547	538	607	521

3. How many events per device type per day? □ **DATA NO AVAILABLE (?)**

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
ios							
android							
Desktop Web							
Mobile Web							

4. How many events per page type per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Search Page	3,995	7,219	7,307	7,221	6,979	7,201	7,137
Book Page	1,977	3,548	3,576	3,572	3,586	3,424	3,506
Driver Page	965	1,823	1,871	1,794	1,755	1,689	1,768
Splash Page	2,954	5,466	5,488	5,376	5,280	5,380	5,184

5. How many events for each location per day?

Date	10/5/2019	10/6/2019	10/7/2019	10/8/2019	10/9/2019	10/10/2019	10/11/2019
Manhattan	6,869	12,591	12,807	12,180	12,270	12,371	12,201
Brooklyn	2,009	3,737	3,590	4,025	3,440	3,400	3,556
Bronx	250	533	507	469	510	394	558
Queens	595	842	905	893	1,026	1,069	936
Staten Island	168	353	393	396	354	460	344

ETL Automation and Scalability:
Provide an analysis about this ETL process. Address and provide rationale for manually extracting, loading and transforming the data from the raw logs. Also address potential preliminary recommendations on improving this process.

The extraction of the data was made directly from the database of the company, which registeres the data, hence, the row data was extracted. After that, some transformations were made in order to format the date column and in order to separate year, month and day, each in a column.
When loading the data, Tableau Public was used, so exploring is more visual and intuitive.

Section 4: Choosing Relevant Dataset

The previous exercise gave you a sneak peek into the Extraction and Loading aspects of ETLs in data pipelines. For making business decisions, a data consumer would like to have all the data they want. However, for any ecosystem, it is impossible to collect or provide everything that the customers need. In this exercise, you will get a taste of real world scenarios wherein:

- All the resources are not always available to get what you need.
- You have to get creative and get the most insights with a minimal data set.

Oftentimes your stakeholders/customers will “ask for the moon”, but you’ll have to push them to work with the small amount of information you have and get creative.

Note: As you learned in the course, being a Data Project Manager involves an extraordinary amount of collaboration. Complete the next sections based on the following scenario.

After the analysis in section 3, we made sense of the numbers, and realized the total number of events seems to be too small (this was a week's worth of data, but you need at least a month). Further investigation reveals that this was a subset of logs, but the actual data that is being collected is much bigger. Working through this small data set was tedious, and repeating this exercise on a much bigger data set manually won’t be feasible. Considering the time constraints of this project, engineering is willing to help with some automation. They also have limited bandwidth and are busy scaling systems up.

Engineering is willing to provide some data, but they have asked for the criterion that is most important. To First provide your business question and provide a rationale for why this is the most important.

Choose one of the following prompts that you think can get you the most relevant information to proceed further.

1. How many events are being recorded per day?
2. **How many events of each event type per day?**
3. How many events per device type per day?
4. How many events per page type per day?
5. How many events for each location per day?

For your chosen question also answer the following using the data from section 3 to support your answer:

1. How much is the customer data increasing?
2. How much is the transactional data increasing?
3. How much is the event log data increasing?

Which of the following data is ***most*** important to answer this question? Why?

- Event Log Data
- Transactional Data
- Customer Data

Answers to the questions above

If I would need a relevant information to proceed further, among the offered options, I would choose "How many events of each event type per day?". With this information it is possible to identify which event type is a pain point for the products success and when this problem takes place. Even though, the other offered information would also be valuable, the event type information is critical.

So, for the "How many events of each event type per day?" question, following questions will be addressed:

It is worth to mention that the data on the 12th of October the data is only available until 10 a.m. This is why for this specific day there is a significant drop for all the three questions below.

1. How much is the customer data increasing?

Date	amount_users	diff_user (%)
05/10/2019	1,644	
06/10/2019	2,743	66.85%
07/10/2019	2,746	0.11%
08/10/2019	2,735	-0.40%
09/10/2019	2,723	-0.44%
10/10/2019	2,692	-1.14%
11/10/2019	2,724	1.19%
12/10/2019	1,382	-49.27%

2. How much is the transactional data increasing?

Date	user_count	diff. user_count (%)
05/10/2019	38.00	
06/10/2019	49.00	28.95%
07/10/2019	62.00	26.53%
08/10/2019	86.00	38.71%
09/10/2019	57.00	-33.72%
10/10/2019	57.00	0.00%
11/10/2019	78.00	36.84%
12/10/2019	18.00	-76.92%

3. How much is the event log data increasing?

Event Date	user_count	diff. user_count (%)
05/10/2019	9,891	
06/10/2019	18,056	82.55%
07/10/2019	18,202	0.81%
08/10/2019	17,963	-1.31%
09/10/2019	17,600	-2.02%
10/10/2019	17,694	0.53%
11/10/2019	17,595	-0.56%
12/10/2019	7,979	-54.65%

Section 5: Loading and Visualization On Your Own

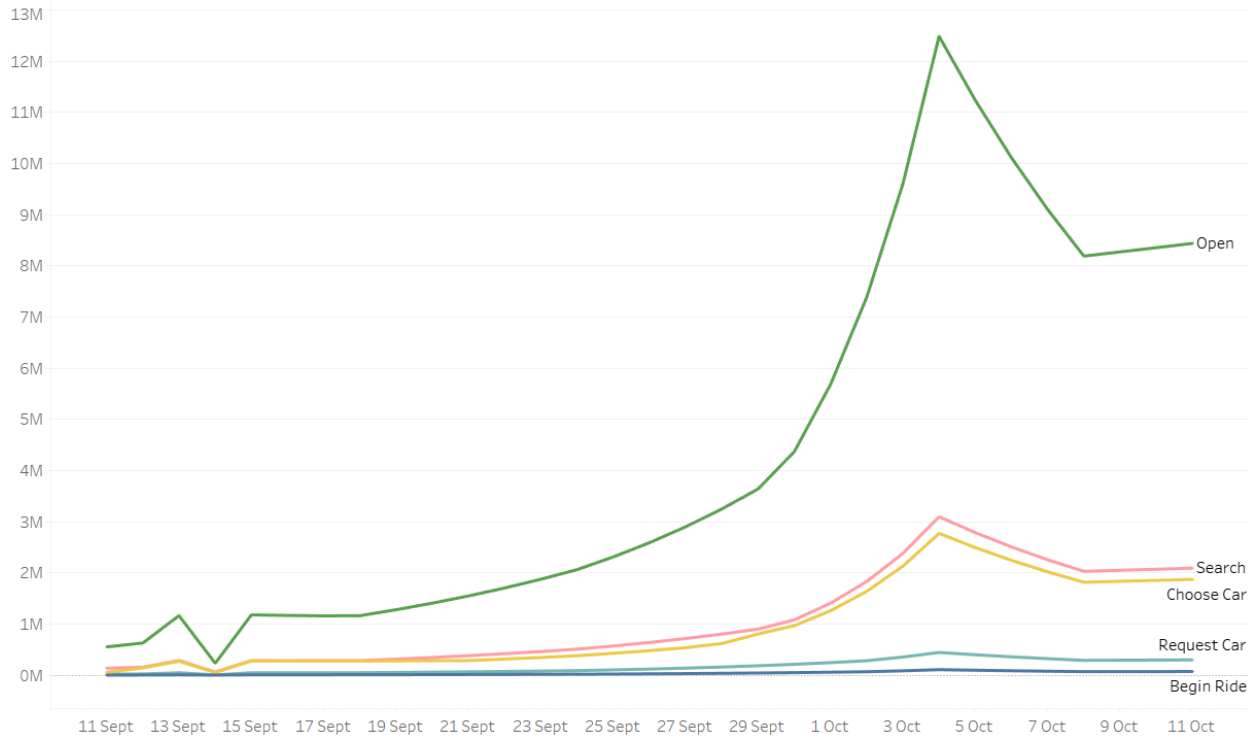
This sectional is an optional part of the project that you can do to make it standout. We have provided visualizations in the appendix if you decide not to do this section. You can also use our visualizations to compare what you created

After sharing your criterion with engineering, they give you a new set of data: Section 5 Event Type Log also available in the classroom resources. Also provided in the project resources section.

Engineering provided you with the data you want, but you still have yet to achieve your ultimate goal as a Data Product Manager. Now, utilize the data to make business decisions. Your executives do not want you to give them a bunch of data tables; instead, they prefer visualizations to help convey the key insights succinctly. Visualizing this data will help you understand the underlying trends and help you determine the story that needs to be told in your proposal to executives.

In this section, you can load and visualize the data into whatever platform you would like. A Python Notebook, Tableau or any other visualization tool you are familiar with. Create two visualizations that might help you to better understand your data trends and place either a screenshot or exported image of your visualizations and the details of each below. Please provide the steps you took to visualize your data and what the visualization tells you about your data.

Visualization 1:



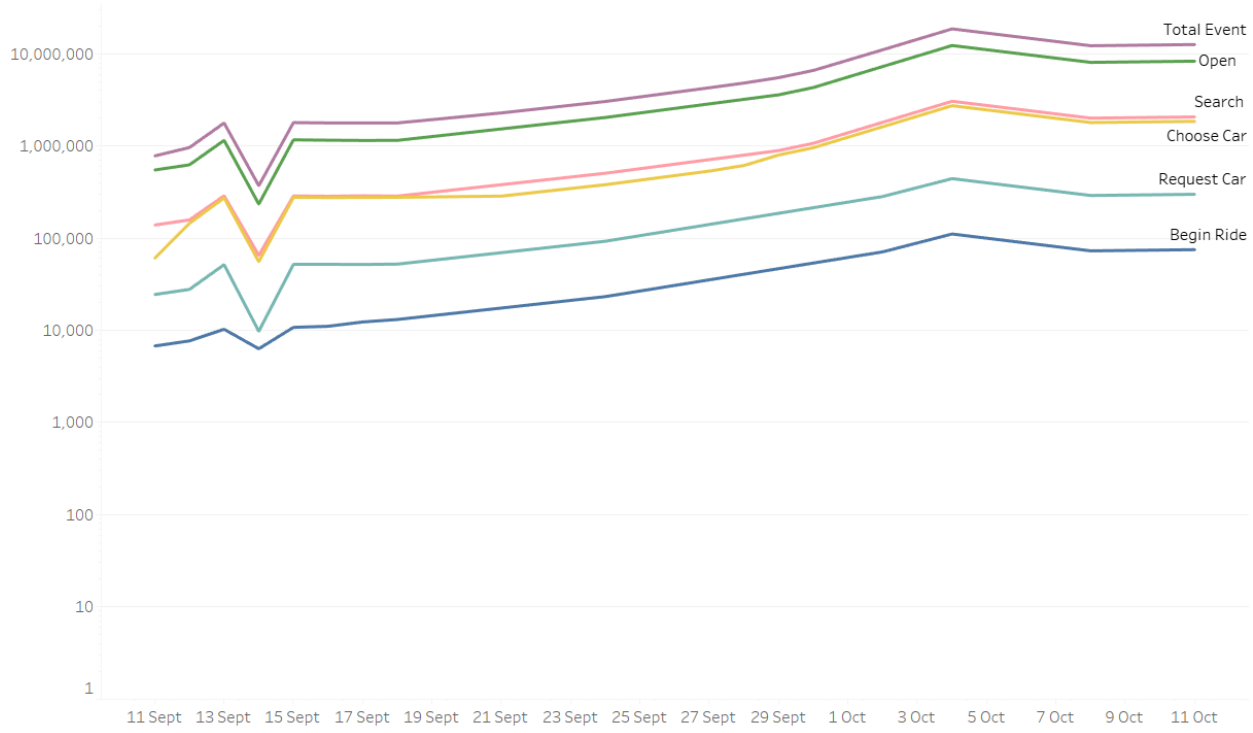
Data Story: This graph tells us:

In the line chart above we can see the amount of observations for each event type. As the process moves forwards the amount of the events for each stage decreases, being "Begin Ride" the one with least amount of observations, meaning this amount of customer purchases. This gives us an overview about which stage is the one that entails the highest amount of observations with drop offs, being this the step from "Open" to "Search".

This graph was created using the following steps:

1. *Connect the data source to Tableau.*
2. *Place the dates into the columns in Tableau and the event type into rows.*
3. *Choose the "line chart" option.*
4. *Distinguish each event by color.*

Visualization 2:



Data Story: This graph tells us:

This visualization provides us the same information as the visualization 1 but using the logarithmic scale in the y-axis. The "Total Event" field was also added, being this the total sum of the rest of the events.

This graph was created using the following steps:

1. *Connect the data source to Tableau.*
2. *Place the dates into the columns in Tableau and the event type into rows.*
3. *Add the "total" field to the rows.*
4. *Choose the "line chart" option.*
5. *Choose logarithmic scale for y-axis.*
6. *Distinguish each event by color.*

	2019	
	September	October
Open	36,255,907	98,860,617
Search	9,013,354	24,542,712
Choose Car	7,438,182	21,971,983
Request Car	1,719,051	3,630,812
Begin Ride	428,490	910,102

Data Story: This graph tells us:

In this graph we can see the number of events by type separated by month. This way, we can identify in which stage do the highest drop offs take place.

This graph was created using the following steps:

1. *Connect the data source to Tableau.*
2. *Place the dates into the columns in Tableau and the event type into rows.*
3. *Choose the "table" option.*
4. *Count the amount of events.*

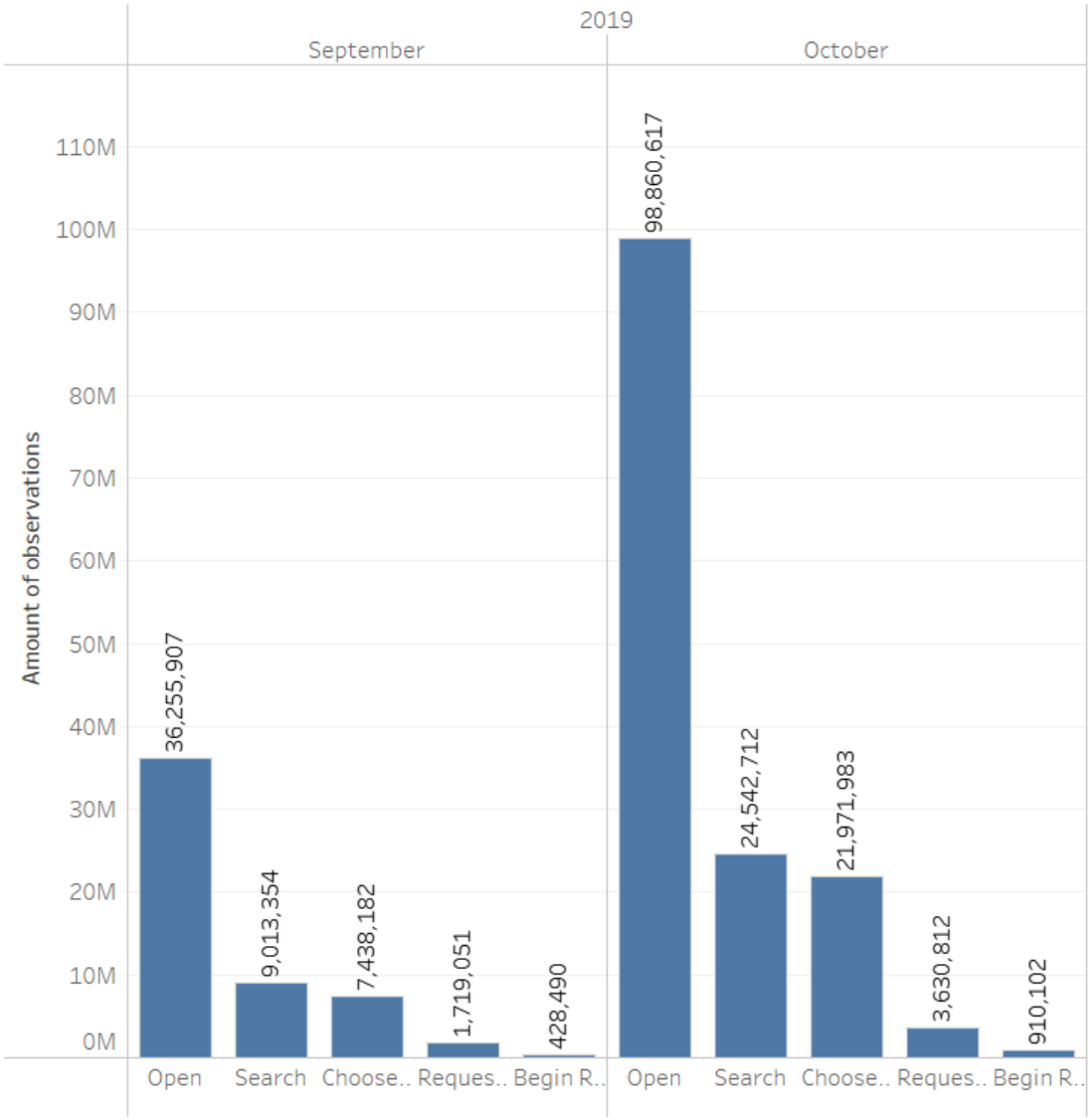
Section 6: Business Insights

The Data is loaded and ready for analysis. We want to use this data as evidence to support our recommendations. It is important that we understand this data and the underlying trends and nuances that these visualizations show us. As you already know, any proposal backed up by data is always better received and considered more robust.

What is the story the data is telling you about Flyber's data growth? If you created Visualizations, you can use them as well, but they are not required). Include any data and calculations that were made to help tell that story and quantify the data growth.

Data Growth for Last Month

Visualization:



Data and calculations used for quantifying of Flyber's Data Growth:

- *Used data: amount of events per stage by month.*
- *Comparison of number of events by stage:*
    - *Open (98,860,617 / 36,255,907) *100 = 272.7%*
    - *Search (24,542,712 / 9,013,354) *100 = 272.3%*
    - ***Choose car*** *(21,971,983 / 7,438,182) *100 =* ***295.4%***
    - *Request car (3,630,812 / 1,719,051) *100 = 211.2%*
    - *Begin ride (910,102 / 428,490) *100= 212.4%*

What is the fastest growing data and why?

The fastest growing data is the one in the stage "Choose car" because the increase in percentage is the greatest one (see calculations on the page before).

What is the Data Story our data tells for each of the following:

- Graph Pattern
- Good or Bad
- October Marketing Campaign
- Marketing Campaign Impact
- Importance of Relationship Between Marketing Campaigns and Data Generation

As we can see in the bar chart above, the data has increased for all the stages, including the purchasing step, which is really good and positive for the business. A big part of the increase can be directly addressed to the marketing campaign that took place in October and which, according to the results shown, has been highly effective. This increase in data generation provides us the information about the importance that marketing campaigns have and their direct relationship with data generation.

Section 7: Data Infrastructure Strategy

Thus far we have:

- identified data stakeholders and their data needs.
- Identified what data is currently being collected and what data needs to be collected.
- Identified data insights and growth trends.

Now, it's time to tie all the loose threads together and bring this process to its logical conclusion by suggesting which Data Warehouse (DWH) Flyber should invest in and why. Using data warehouse options below, suggest whether Flyber should choose an on-premise or Cloud data warehouse system and which specific data warehouse would best serve Flyber's data needs.

Data Warehouse Options:

Cloud:

- Amazon Redshift
- Google BigQuery
- Snowflake
- Microsoft Azure

On-Premise:

- Oracle Exadata
- Teradata, Vertica
- Apache
- Hadoop

You will address the following factors with a rationale as to why the DWH chosen is the best for Flyber:

- Cost
- Scalability
- In-house Expertise
- Latency/Connectivity
- Reliability

Cloud vs On-Premise

Provide an evidence based solution as to why Flyber would be best served by a Cloud or on-premise DWH. In this response, you don't need to specify *which* specific Cloud or on-premise DWH product you will choose, just if it will be Cloud or on-premise. Remember to address the factors above.

We would choose a cloud DWH for the following reasons:

- The infrastructure costs will be lower in the long term. Additionally the costs will be more predictable to defined software license cost, support, etc. Another reason is that the maintenance costs are included.
- As we saw in the Section 6 the increase of data is being really fast so Flyber will have the need of scalability, which is a really easy to get with cloud services by adding licenses.
- Regarding the in-house expertise, as Flyber is a new company that is growing fast there is not high expertise in DWH so using a cloud solution will facilitate the implementation, maintenance and use.
- The DWH in Cloud will be available from everywhere and the latency can be scaled depending on the need.
- One con of cloud solutions is the lower control over security and compliance.

Suggested DWH

Provide an evidence based solution as to which DWH (Data Warehouse) product is best for Flyber. Remember to address the factors above.

Microsoft Azure

It is a well known platform, which will be easy to integrate with other available tools in the market. Furthermore, it will be easier to hire employees that know the technology.
Regarding the billing, Azure offers different plans to store the date and also to compute it. This can be scaled depending on the needs by adding space, licenses or power.
If we focus on the costs, Azure will maintain the BWH, will offer a fixed billing plan and has great customer service. The only concern will be the lower control over security and compliance since Flyber will not have complete control over what is being done behind the scenes.